DOCUMENT RESUME

ED 183 176 IR 008 115

AUTHOR Lowry, Glenn R.
TITLE Online Document Retrieval System Education for Undergraduates: Rationale, Content and Observations.
PUB DATE 25 Mar 80
NOTE 17p.; Paper presented at the National Online Information Meeting (New York, NY, March 25, 1980)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Libraries; Course Descriptions; *Information Retrieval; Information Systems; *Library Instruction; *Library Skills; *On Line Systems; Undergraduate Students
IDENTIFIERS *New Jersey

ABSTRACT

This report describes the conceptual content and hands-on activities of a course in online information retrieval skills offered at Stockton State College to expose undergraduate students to an important area of information science not traditionally included in undergraduate computer science curricula. Observations of student experience with the course work, an evaluative summary, and references are provided. (FM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183176

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Online Document Retrieval System Education for Undergraduates: Rationale, Content and Observations

by

Glenn R. Lowry

Faculty of Professional Studies
Stockton State College

A paper to be delivered
at the National Online Information Meeting,
March 25, 1980.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Glenn R. Lowry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR008115

Online Document Retrieval System Education for Undergraduates

by

Glenn R. Lowry

Abstract

A course in basic information retrieval principles and use of online document retrieval systems is a curriculum requirement for undergraduate computer science students at Stockton State College in New Jersey. A combination of theory-oriented lectures and online search sessions using DIALOG enables students to observe course principles in action. An undergraduate course of this type differs significantly in content from those normally offered in graduate library schools in one primary area; while most students in graduate library schools are presumably aware of the functions and issues of index languages and library operations, undergraduate computer science students need to be taught the basics of these subjects. A corequisite course insures student background in the design requirements of deterministic data retrieval systems. This background provides a useful framework in which to explore index lanugage limitations and design features of document retrieval systems which must provide multiple access points. The course has been enthusiastically received as evidenced by student participation, anonymous student evaluations and requests for the development of subsequent courses designed to prepare students for careers in the online information retrieval service environment.

Introduction

Proliferation of online document retrieval services, increased public awareness of their availability and sustained demonstrations of their economic viability suggest a pattern of growth which is likely to continue in the foreseeable future. Acknowledging the permanence of these services in library practice, Bellardo, Kennedy and Tremoulet (1) detailed a thorough and somewhat original approach to online bibliographic system instruction for graduate library students. While numerous approaches to online education are available, as reported by Harter (2), most are geared to graduate students in schools of library and information science. As automated reference services continue to gain in importance and number, it is likely that library professionals will begin sharing at least some of the technical routines of online bibliographic retrieval with non-professionals. At the same time, organizations which design and operate the systems are likely to require increasingly greater numbers of computer scientists who have been exposed to principles of information retrieval. Computer scientists should be brought to an awareness of desirable system performance features as a part of their education. In response to projected staffing needs in these areas, a one-semester course in information retrieval and technology is included in the required curriculum of study for undergraduate information and systems sciences students at Stockton State College.

Prior to course development, some general objectives were formulated into an operating course rationale. The course should include practical, live instruction in Boolean literature searching techniques employed in commercial online databases. The course should also include study of basic principles of information retrieval, including the objectives, advantages and limitations of current systems. Consideration of index language characteristics and trade-offs, system performance measures and computational considerations should be a part of the course. It was felt that the practical and theoretical elements of the course would mutually support and reinforce one another. It was desired to prepare students for more detailed study of online system design and programming, to stimulate student interest in such study, and to instill in students an appreciation of the potential for comprehensive bibliographic control and current awareness support available through online interactive information retrieval systems.

Due to the limitations of time the course could not include extensive work in programming and file design by students, nor could advanced research topics be more than alluded to. Issues relating to artificial intelligence and automatic query modification were not touched.

As a corequisite or prerequisite, students were required to complete a course dealing with business data processing and design of management information systems. Students so prepared should be better able to appreciate design and operational distinctions

between deterministic management information systems and stochastic, interactive online retrieval systems in which file content and output possibilities are not known to the end user in advance. When this distinction has been made, usually after the initial online literature search by the student, the groundwork has been laid for an appreciation of the purposes of indexes, index languages and document surrogation. Students then begin to explore the intellectural challenges, limitations, and possibilities inherent in these processes. In the remainder of this paper, attention will focus on the conceptual content of the course, the extent and character of hands-on experience, some observations of student experience with the course work, and a summary.

Conceptual Content of Course

Many courses in online bibliographic retrieval offered at graduate library schools share several characterisitics in common. First, students enrolled in such courses are likely to be quite knowledgeable in the areas of library operations, classification systems and indexing theory. These same students, however, are typically unschooled in data processing, data representation in computers, or in computer capabilities generally. Considerable effort must be expended in helping the student to acquire a basic understanding in these areas. Once this has been accomplished, the instructor is free to present the substantive areas of the course. These include basic information retrieval principles dealing with index languages, performance measures, and search

strategy. These areas are illustrated through student hands-on experience in Boolean searching in online commercial data bases. This approach, successful in library schools, would not be likely to succeed in an undergraduate computer science program.

While the substantive content of the undergraduate course remains much the same as that of the library school in the areas of theoretical and experiential material, the type of preparation needed by the undergraduate in order to make the substantive content meaningful is quite different from that which the graduate library school student needs. The undergraduate generally has no prior knowledge of or experience in dealing with library-related concepts, classification systems or indexing theory. A good deal of course material is directed toward helping the student to arrive at a basic level of understanding of these subjects. Approximately three weeks of a fifteen-week semester are devoted to this essentially preliminary and remedial activity. This does not, however, preclude effective coverage of substantive course material. A similar period must sometimes be devoted to familiarizing library school students with computers, leaving approximately the same time available for substantive course material in either environment.

The undergraduate course at Stockton focuses on three conceptual areas: library-related content, basic information retrieval principles, and computational considerations.

The course begins with a consideration of the nature and sources of information and attempts to help the student distinguish between the data represented in knowledge-records and information,

which subsists in the minds of human beings. When the student sees the loss of correspondence which necessarily occurs between ideation and communication, he begins to appreciate some of the difficulties encountered by indexers and retrieval system designers. Enumerative and analytico-synthetic classification systems are introduced and the characteristics and limitations of each are compared.

Library operations are discussed, focusing on the purposes, operational methods and problems of cataloging, acquisitions and circulation activities. This prepares the student for subsequent study of index languages and their limitations and trade-offs. It also enables the student to understand problems encountered in subject coverage decisions and overlap between data bases. It prepares the student to connect index term use data with ongoing term definition modification and subject retrieval profiles with acquisition decisions. Through library-related material, students are helped to an understanding of one of the key objectives of an online document retrieval system, that of successfully preparing a document file which will facilitate document retrieval by a client whose specific background, information needs and method of stating those informational needs are unknown in advance.

In preparation for the hands-on portion of the course, students are introduced to DIALOG. They are required to purchase Lockheed's publication, A Brief Guide to DIALOG Searching (3) as one of two required texts. The "library" portion of the course is concluded by instruction in the use of DIALOG query language and the formula-

tion of Boolean search strategies for use online.

After an initial online search, to be discussed below, attention moves to presentation of basic information retrieval principles. A second text, Meadow's The Analysis of Information Systems (4) is the primary reading source supporting this portion of the course. Supplementing the readings from Meadow, relevant sections of Lancaster's Information Retrieval Systems (5), Artandi's An Introduction to Computers in Information Science (6) and Sharp's Information Retrieval (7) are assigned. Readings from important journal articles are also included.

The computer is approached as a communication device, as an instrumentality for conveying messages from authors to recipients through a communication channel. The work of Shannon and Weaver (8) is discussed briefly. A conception of the limitations of mathematical communication theory is developed as relating to the potential of a given channel for syntactic transmission of signs.

Next, index languages are approached, beginning with an introduction to the document surrogation process. The importance of index terms as the end result of the conceptual analysis performed by an indexer is stressed. Non-syntactic and syntactic index languages are discussed in turn. The discussion moves from hierarchical, through subject heading and fixed and free-key word languages to syntactic languages, including tagged descriptors, phrases and natural language. Each of these index language types is contrasted with the others so that the student can gain an appreciation of the strengths, weaknesses and trade-offs which pertain to each.

Following this, attention shifts to statistical selection of index terms, automatic classification, and automatic abstracting. Presentation of these topics is necessarily elementary, beginning with Luhn's classic papers (9) (10) and concluding with readings from Chapter 8 of Borko and Bernier's Abstracting Concepts and Methods (11). The discussion focuses on the importance of term frequencies and co-occurrences in the production of index terms, on the possibilities of machine-determined classification of documents based on comparison of index terms with subject term profiles, and on the production of extracts and abstracts based on index term weighting and clustering techniques.

Retrieval strategies are reviewed, beginning with several examples and in-class exercises in their development. A statement of information needs is predicated. Overhead transparencies are used to demonstrate a way in which concepts within the statement are identified. The student is then shown how to represent each concept by the union of several synonyms and, finally, how the intersection of the sets created by the unions can result in the retrieval of relevant documents. After guiding the student through two or three additional exercises, he is ready to complete one or two of them in class on his own.

Next, the matter of system performance evaluation is discussed. Computed and judged relevance are differentiated. Precision and recall are presented together and it is at this point that the majority of students who have taken the course report that all of the material previously covered begins to make sense and to seem

important to them. Pertinence and utility are also discussed, insuring more than a simplistic, system-oriented conception of user-system interface by students.

The course next turns to computational considerations, an area in which the undergraduates have some depth of background. Here, the discussion begins with record structures. Fixed and variable field-length computer record structures are examined in detail. The discussion includes dealing with repeating fields and with standard length tagged fields. Record structural characteristics such as expressiveness, positional and semantic ambiguity and compactness are explored.

Following this, file structures are introduced. With the aid of numerous transparencies, the detail and differences of a variety of file structures are highlighted. Initially, sequential and inverted file structures are examined and distinctions are made between them. Other file structures are also considered, including chain, branching and list structures.

Finally, file sets are discussed in terms of design purpose. Fixed sequence searching is presented as geared primarily toward the non-specialist end user while variable sequence searching is more appropriate for the trained intermediary who has knowledge of the file structure of the system he is using, of the system's capabilities and of basic information retrieval principles. Sample processing algorithms are discussed briefly throughout this phase of the course which concludes consideration of conceptual and computational material.

Hands-On Experience

A vital part of the course at Stockton is hands-on experience in searching online data bases. By working with a live, commercial data base, information retrieval principles studied theoretically are vividly operationalized for the student. Theory and practice mutually reinforce one another. Each student plans and conducts two online searches in DIALOG. This experience is made possible by Lockheed's generous Classroom Instruction Program.

The first student search session is conducted approximately 20% of the way through the course. The purpose of this initial search session is to familiarize the student with an operational information retrieval system and with use of the system's query language. DIALOG serves as an excellent model, demonstrating how information retrieval concepts have been successfully put into action. This initial search helps the student to internalize information retrieval concepts discussed in the remainder of the course. The search is evaluated liberally, on the basis of familiarity with the command language and system features and not on the actual results of the search. Each student receives personal instruction for both pedagogical and budgetary reasons. The initial search usually lasts for approximately thirty minutes. System features encountered during the initial search are referred to during the conceptual portion of the course to illustrate theoretical concepts as they are discussed.

The course culminates in a final student online search session. This session provides a final opportunity to reinforce course concepts and for the student to demonstrate mastery of subject matter and increased familiarity with the DIALOG system. This search is also personally supervised but is conducted with as little intervention by the instructor as possible. It is evaluated by comparing retrieved records with a search statement formulated by the student prior to going online. Evaluations from the two search sessions comprise approximately one half of the student's grade. The other half is the result of four essay examinations each of which contains three questions. These are drawn from four lists of eight or nine review questions with which the student is provided at least a week in advance of each examination. The study questions help to reduce feelings of bewilderment expressed by students during the first half of the course and help to focus their thought on important course material. Most of the questions require library work to answer fully. The nature and number of questions confronting the student promotes teamwork as evidenced by the formation of informal "study groups".

Observations

Undergraduate computer science students learn Boolean search strategy and DIALOG query language readily. They enjoy working with the system as evidenced by the high number of requests for additional online search time received by present and former students. While the only persons required to remain at the terminal during a given class period are those scheduled for search sessions,

most class members typically attend all sessions and tend to remain at the search facility long after the search period has concluded.

Students consistently report that the hands-on experience gained during their two search sessions did succeed in providing them with a model to which to relate concepts studied in the remainder of the course.

Undergraduate computer science students at Stockton State College are normally goal-oriented and highly motivated due perhaps to good job prospects and to the high prestige which they enjoy among their peers. They are skills-oriented, an extension, perhaps, of their career objectives. They are curious and inquisitive and usually ask many questions. These same students, often familiar with the instructor from previous courses in programming or systems analysis, are extremely quiet, almost non-responsive for approximately the first half of the course. During this period, they write furiously and frequently visit the instructor to express fears of being lost and feelings of bewilderment at not knowing how course material connects. Each time the course has been offered, the turning point has been the discussion of relevance, precision and recall. Many students have reported that, following that particular discussion, they were able to integrate preceding course material into a coherent body of knowledge. In subsequent conversations, former students report that their confusion up to that crucial point is due to the newness to them of much of the course matter rather than boredom with that material or with the instructor. Although no empirical research has been done to establish the

generality of these observations, some other factors seem to support this view. Attendance in class is not mandatory but is uniformly high. In addition, numerous students have expressed interest in development of a subsequent course dealing with design and programming of interactive document retrieval systems. Some have expressed interest in employment in the online information retrieval industry. Finally, in anonymous student evaluations of "overall course effectiveness," the course received high composite effectiveness ratings each time it has been offered. Clearly, those who have taken the course feel that it is of considerable value to them.

Summary

The present course in information retrieval at Stockton State College is an attempt to expose undergraduate students to an important area of information science not traditionally included in undergraduate computer science curricula. By providing an enriched mixture of theory-oriented lectures and online search sessions, the course is intended to stimulate student interest and enthusiasm for further study of online information systems. At the very least, a course objective is to confirm students as positive users of online information retrieval services. As the course is currently in its fourth term, it is not possible at this time to assess longer term effects which may develop in the careers of former students.

As the online information retrieval industry continues to grow, as the numbers and magnitudes of data bases increase, it is not unlikely that the need for computer scientists trained at the baccalaureate level to be familiar with the principles of information retrieval will increase accordingly. This may promote more attention to the need for training in this area in increasing numbers of computer science departments and by their students in response to attractive employment opportunities. Investigation of this possibility is planned.

In times of ever-shrinking budgets, hungry library administrators may begin to view para-professionals trained at the undergraduate level in a course such as this as an economical alternative to a professional who commands a higher salary. While this possibility could negatively impact upon future employment of library professionals hired to conduct online literature searches, it may be ultimately beneficial in freeing professionals for more involvement in information counselling, helping to serve more clients more effectively. Performance comparisons between undergraduate computer science students and graduate library school students as online search intermediaries are also planned.

Whatever the outcome of these and other issues confronting us at present, the next several decades will no doubt prove extremely stimulating to those fortunate enough to be involved in some way with online information retrieval systems.

REFERENCES

1. Bellardo, Trudi; Kennedy, Gail; and Tremoulet, Gretchen. "On-line bibliographic system instruction," Journal of Education for Librarianship. V.19 (Summer 1978), pp. 21-31.

2. Harter, S. P. "Instruction provided to library schools in machine-readable bibliographic data bases," Proceedings of the ASIS Annual Meeting 14:49, 1972.

3. A Brief Guide to DIALOG Searching, 2nd revised edition. Palo Alto, California: Lockheed Missiles and Space Co., Inc., 1978.

4. Meadow, Charles T. The Analysis of Information Systems, 2nd. edition. New York: John Wiley & Sons, 1973.

5. Lancaster, F. Wilfrid. Information Retrieval Systems: Characteristics, Testing and Evaluation, 2nd. edition. New York: John Wiley and Sons, 1979.

6. Artandi, Susan. An Introduction to Computers in Information Science, 2nd edition. Metuchen, N.J.: Scarecrow Press, 1972.

7. Sharp, John R. Information Retrieval: Notes for Students. London: Andre Deutsch, 1970.

8. Shannon, Claude E., and Weaver, Warren. The Mathematical Theory of Communication. Urbana, Illinois: University of Illinois Press, 1959.

9. Luhn, H.P. "The automatic creation of literature abstracts," IBM Journal of Research and Development. Vol. 2(1958), pp. 159-165.

10. Luhn, H.P. "A statistical approach to mechanized encoding and searching of literary information," IBM Journal of Research and Development. Vol. 1(1957), pp. 309-317.

11. Borko, Harold and Bernier, Charles L. Abstracting Concepts and Methods. New York: Academic Press, 1975.